Dialogue & Discourse 12(2) (2021) 145–173 doi: 10.5210/dad.2021.205

Lexical Alignment to Non-native Speakers

Iva Ivanova IMIVANOVA@UTEP.EDU
Department of Psychology
University of Texas at El Paso
500 W. University Ave., El Paso, TX 79902, USA

Holly P. Branigan HOLLY.BRANIGAN@ED.AC.UK
Department of Psychology
University of Edinburgh, Edinburgh, UK

Janet McLean J.MCLEAN@ABERTAY.AC.UK
Division of Psychology
Abertay University, Dundee, UK

Albert Costa SECRETARIA.DTIC@UPF.EDU
Departament de Tecnologies de la Informació i les Comunicacions
Universitat Pompeu Fabra, Barcelona, Spain
Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Martin J. Pickering MARTIN.PICKERING@ED.AC.UK
Department of Psychology
University of Edinburgh, Edinburgh, UK

Editor: Patrick G. T. Healey

Submitted 02/2019; Accepted 04/2021; Published online 10/2021

Abstract

Two picture-matching-game experiments investigated if lexical-referential alignment to non-native speakers is enhanced by a desire to aid communicative success (by saying something the conversation partner can certainly understand), a form of audience design. In Experiment 1, a group of native speakers of British English that was not given evidence of their conversation partners' picture-matching performance showed more alignment to non-native than to native speakers, while another group that was given such evidence aligned equivalently to the two types of speaker. Experiment 2, conducted with speakers of Castilian Spanish, replicated the greater alignment to non-native than native speakers without feedback. However, Experiment 2 also showed that production of grammatical errors by the confederate produced no additional increase of alignment even though making errors suggests lower communicative competence. We suggest that this pattern is consistent with another collaborative strategy, the desire to model correct usage. Together, these results support a role for audience design in alignment to non-native speakers in structured task-based dialogue, but one that is strategically deployed only when deemed necessary.

Keywords: audience design, communicative success, lexical choice, picture-matching game

1. Introduction

How we talk may feel unique but is not always original. We sometimes mimic aspects of the utterances of our conversation partners, a phenomenon known as alignment (Pickering and Garrod,

2004) . Alignment is beneficial for conversations in that it facilitates mutual understanding (Ferreira et al., 2012). Speakers in dialogue can align to each other's words (*lexical alignment*; Brennan and Clark, 1996), phrasal or sentence structure (*structural alignment*; Branigan et al., 2000), as well as phonetic or prosodic aspects of utterances (Giles and Powesland, 1975; Pardo, 2006). However, alignment and its underlying cognitive mechanisms may vary with features of the conversational situation and characteristics of the conversation partner. One such feature is interacting with a non-native speaker. Here, we study whether lexical alignment with non-native speakers is influenced by a desire to aid a conversation partner's comprehension, a form of audience design. We address this question in two experiments involving a constrained exchange of referential expressions between participants and confederates, to be able to implement specific manipulations that test our hypotheses. At the end, we discuss differences with, and implications for, less constrained natural dialogue.

A classical experimental demonstration of lexical alignment (and ensuing alignment of situation models) was provided by Garrod and Anderson (1987). In a cooperative maze game, one participant had to indicate her or his position in a maze to another participant viewing the same maze from another room. Participants tended to converge in their descriptions of the maze structure and their positions (for example, Participant A: "Right - *two along from the bottom* one up"; Participant B: "*Two along from the bottom*, which side?"). Further demonstrations have involved tasks requiring one participant to relate the order of pictures of objects that can be named in more than one way to another participant who had to arrange her or his figures or images in the same order Brennan and Clark (1996). In such tasks, participants in a dyad typically adopt the same way of referring to the figures or images (for example, if one participant refers to a picture of a shoe as "pennyloafer", the other participant is more likely to use the same term to refer to it).

Lexical alignment may be important for communicative success. First, it may aid mutual understanding, by helping conversation partners reach a similar mental representation of a situation. Ferreira et al. (2012) showed that participants were faster to choose pictures on a display when their own names for those pictures were repeated back to them versus when other words were used. Second, imitating the behavior of one's conversation partner (including reusing the words they produce) might increase rapport, empathy and prosociality between the conversation partners (for a review, see Beňuš, 2014; Chartrand and van Baaren, 2009; Garrod et al., 2018). In addition, alignment may underlie long-term learning: Reusing a language representation (including a word representation) may strengthen its connections to other elements in the language system and thus facilitate its reuse on future occasions (see Oppenheim et al., 2010; Chang et al., 2006, for structural alignment). Of note, one study found that lexical alignment conferred communicative benefits only when it was limited to task-relevant vocabulary, suggesting that its effects are well-targeted; overall alignment was in fact negatively correlated with task performance (Fusaroli et al., 2012).

The potential of lexical alignment to aid communication might make it an especially important communicative mechanism in conversations with non-native speakers because these speakers face various challenges in conversing in a second language. Speaking a second language is cognitively demanding and associated with reduced fluency and clarity (presumably stemming from word-finding difficulties; Pivneva et al., 2012) and delayed lexical access (Ivanova and Costa, 2008). Understanding a second language is also cognitively effortful and deteriorates in the presence of noise (for the latter, see Weiss and Dempsey, 2008). Still, an increasing number of people converse in a second language with native speakers of that language, for purposes of work, education or

travel - but little is known about the functioning of alignment in such conversations (see Costa et al., 2008).

To begin understanding lexical alignment with non-native speakers, it is necessary to consider the cognitive mechanisms that drive it. One such mechanism is audience design: speakers' tendency to construct their utterances considering the knowledge, needs and intentions of their conversation partners (Brennan and Clark, 1996; Clark and Schaefer, 1987). Audience design is goal-driven and involves belief-based judgments about what would be more versus less effective for conversation partners' comprehension (Clark, 1996), or what would be maximally informative (Grice, 1975). Lexical-referential alignment is among the strategies that ensure successful information transmission: The use of a particular referential expression by the conversation partner is clear evidence that they will understand it if repeated back to them. Indeed, comprehension is slowed down when listeners hear different terms than they themselves (Ferreira et al., 2012) and also their conversation partners (Metzing and Brennan, 2003) have previously produced in the course of their interaction.

Several studies support that audience design can drive lexical alignment. These studies have shown that speakers can override their normal production preferences and adopt the lexical choice of their conversation partners based on their judgment of the partners' likelihood of experiencing comprehension difficulties. In a director-matcher task in which participants could interact freely, Isaacs and Clark (1987) showed that New-Yorkers (experts) described postcards of buildings in New York City to non-New-Yorkers (novices) using fewer building names (which the novices would not know) than to other New-Yorkers (and decreased their use of proper names with non-New-Yorkers by approximately 20% across trials). This result supports audience design by showing that experts adapt their lexical choices to novices' productions, and presumably do so to ensure that their linguistic contributions will be understood.

Most relevant to the present study, Branigan et al. (2011) showed that speakers aligned their lexical choice to computers more than to humans, and to computers perceived as older more than to computers perceived as newer. In this study, participants alternated between describing pictures either orally or in writing to what they were told was a human or computer conversation partner, and matching pictures to the "partner's" picture descriptions (in reality, all such descriptions were scripted). On experimental trials, the "partner" produced either preferred (for example, *bus* for a picture of a bus) or dispreferred but acceptable responses (for example, *coach* for a picture of a bus). There was greater lexical alignment with conversation partners presumably perceived as less capable of producing accurate descriptions. The authors concluded that this tendency was driven by assumptions about what the conversation partner could versus could not understand (that is the words they used versus words they did not use) instead of a desire for a social bond. This is because people would arguably not want to create a social bond with a computer to a greater extent than with another person.

Both of these studies suggest that a belief that a conversation partner is less communicatively competent (a "novice" or a computer) can increase lexical alignment, supporting a role of audience design for alignment. In the context of the current study, audience design might also act to increase lexical alignment to non-native relative to native speakers, to the extent that non-native speakers can be perceived as less communicatively competent conversation partners because of their incomplete knowledge and slower processing of the target language. Further, audience-design-driven lexical alignment to non-native speakers may be inversely proportional to how communicatively competent they seem, and hence increase when beliefs or perceptions point to lower competence (consistent with the larger alignment with "older" computers in Branigan et al., 2011).

Strong evidence about what a conversation partner does or does not understand is feedback that provides *grounding* of a speaker's utterances – that is, contributes to the mutual belief of sufficient understanding of the entities under discussion to move forward (Clark and Brennan, 1991; Clark and Schaefer, 1987; Clark and Wilkes-Gibbs, 1986). In natural conversation, grounding is accomplished in one of three main ways: by providing acknowledgements through backchannel responses (for example, *uh-huh* or *mhm*) or assessments (for example, *Really??* or *Good God!*), by directly initiating the relevant next turn (for example, answering a question in a relevant way), and by continued attention on the speaker as shown through eye-gaze (although the latter can have other functions; Clark and Brennan, 1991). Grounding of referring expressions can also be indicated by alternative descriptions (for example, A: *It is the second one on the left*, B: *You mean the red one?*) or indicative gestures (for example, pointing) to verify understanding.

However, grounding that is optimal for specific conversational needs is not always possible. For example, grounding is affected by the conversational medium. For example, participants in video conferencing often turn off their microphones or cameras, and by doing so they also eliminate the possibility of providing backchannel responses. But even face-to-face conversations do not always provide opportunities for listeners to assert that they understand every detail of what is being said, such as when the speaker is telling a story, or a professor is lecturing to a class. This is relevant in the current context for the following reason. If lexical-referential alignment is used as a tool to ensure successful comprehension, of which feedback of listener understanding, when it occurs, provides direct evidence, alignment may be used as an audience design strategy to a greater extent (or only) when the possibility for feedback is reduced or absent.

Consistent with this possibility, Bergmann et al. (2015) found higher lexical alignment when direct evidence of conversation partner understanding was scarce. These authors showed more lexical alignment with an assumed artificial agent than with an assumed human (in fact both were artificial agents), but only when participants could not see (on video) their conversation partner. These results are consistent with alignment driven by an intention to aid communication because speakers adapted their lexical choice to a greater extent to a partner they judged as less communicatively-capable (such as a virtual agent as opposed to a human). But they also imply that visual feedback of the conversation partner's behavior can eliminate this difference, presumably because being able to see one's conversation partner gives greater certainty of comprehension success (reducing the need for lexical alignment to ensure this success). If so, it is possible that audience design would not increase alignment to non-native speakers overall, but only in the absence of evidence that they correctly understand what is being said.

This possibility is consistent with evidence that speakers do not always consider conversation partners' knowledge and needs because making such considerations is cognitively effortful. Accordingly, Horton and Keysar (1996) showed that participants were less likely to take common ground into account in a referential communication task under time pressure than under no time pressure. The authors concluded that such audience design does not form a core part of designing utterances but is only part of an optional monitoring process, which we assume might be invoked when comprehension success is uncertain.

We note that audience design, our current focus, is not the only mechanism that may drive lexical alignment. Another mechanism that may underlie all alignment behavior is priming of the underlying linguistic representations (Pickering and Garrod, 2004). For lexical alignment, when one's conversation partner produces a particular referential expression (such as the name *sofa* for an object that can also be called *couch*), this name becomes highly activated. This increased ac-

tivation would in itself make it more likely that that name be selected for production over other possible choices. Consistent with a priming mechanism contributing to alignment is evidence for similar lexical alignment in typically-developing children and children with an Autistic Spectrum Disorder, whose theory-of-mind abilities or social functioning did not predict alignment magnitudes (Branigan et al., 2016; Hopkins et al., 2017). In the current study, we focus on the contribution of audience design to alignment to non-native speakers, but assume that priming underlies alignment to both non-native and native speakers at least in part (although possibly to different degrees; we return to this issue in the General Discussion).

A direct examination of alignment to non-native speakers was provided by Bortfeld and Brennan (1997). In a referential communication task, pairs of participants took six turns to be directors (describing a set of 15 chairs to the other person) and matchers (identifying each chair and placing it in the correct serial position). Results showed that native speakers lexically aligned to non-native speakers to the same extent as to other native speakers (even if this involved alignment on non-nativelike expressions, such as *the chair in which I can shake my body* for a rocking chair). But this study did not investigate the relative strength of different mechanisms driving alignment with non-native speakers. The fact that native speakers aligned on non-nativelike expressions suggests an influence of audience design, as they would not have had any other motivation to produce such expressions except to ensure that their non-native partners understood them. On the other hand, non-nativelike expressions would not have direct representations in native speakers' mental lexical inventories and thus would be unable to automatically prime native speakers' productions. Thus, alignment to non-native speakers in Bortfeld and Brennan's study could have been greater (possibly greater than that to native speakers) had the non-native speakers produced more nativelike expressions.

1.1 The Present Study

In the present study, we compared the magnitude of native speakers' lexical-referential alignment to non-native speakers with that to native speakers, to test several hypotheses about the influence of audience design on alignment to non-native speakers in structured task-based dialogue. Hypothesis 1 (tested in both experiments) is that native speakers would align more (and not less) to non-native speakers than to other native speakers, at least under certain conditions. This would be because native speakers would take greater care to ensure comprehension success with non-native speakers because they perceive the latter as less communicatively competent. Hypothesis 2 (tested in Experiment 1) is that such enhanced alignment to non-native speakers would be more, or only, present in conditions lacking feedback of comprehension success (that is, grounding). This would be because direct evidence that the conversation partner is understanding everything would alleviate the need for enhanced alignment with a less competent conversation partner to ensure communicative success. Hypothesis 3 (tested in Experiment 2) is that enhanced alignment to non-native speakers would be modulated by how communicatively competent they appear to be. This would be because enhanced lexical alignment as a strategy to ensure comprehension success may appear more necessary when the non-native conversation partner appears less communicatively competent.

To determine the extent to which patterns of alignment with non-native speakers might generalize beyond particular populations or speakers of a particular language, we investigated two different populations (speakers of British English tested in Edinburgh in Experiment 1, and speakers of Castilian Spanish tested in Barcelona in Experiment 2).

In a picture-matching game, participants took turns with confederates to name pictures for each other and match pictures after hearing the names produced by their partners. Each native participant performed rounds of the game with both a native and a non-native confederate. The confederates (on half of the critical trials in Experiment 1 and on all critical trials in Experiment 2) produced dispreferred names (for example, calling a lamp *light*, or a basket, *hamper*, based on pretested norms; Branigan et al., 2011). We focused on participants' use of dispreferred names to detect differences in alignment across conditions, given that we expect speakers to use preferred names as their default choice (that is, independently of the identity of the conversation partner who produced them previously). In both experiments, each participant interacted with two confederates, in counterbalanced order.

In Experiment 1, native speakers of British English alternated with a native and a non-native confederate to name one of two pictures presented on a printed page, and indicate which of the two pictures on the following page matched the name produced by the other person. In addition, Experiment 1 manipulated the presence of feedback of comprehension success. For this purpose, half of the naïve participants received feedback from the confederate about their communicative success (who verbally indicated the correct response with the words *left* and *right*; participants did the same on their own matching turns). The other half of the participants matched pictures without receiving or providing any such evidence of comprehension success. In Experiment 1, Hypothesis 1 predicts more lexical alignment with the non-native than with the native confederate. Hypothesis 2 predicts that such an effect would be larger (or only present) without feedback.

In Experiment 2, native speakers of Castilian Spanish alternated with a native and a non-native confederate to consecutively name 12 pictures presented on three-by-four grids, and select the picture named by the other person among an initial set of 12 loose pictures (decreasing after each trial), to then place it in its respective position on a blank three-by-four template with numbered grids. In addition, Experiment 2 manipulated the apparent communicative competence of the non-native confederates. Thus for one group of naïve participants, the non-native confederate produced gender errors (for example, "*un mano" instead of "una mano" [Sp. a hand]) on half of the trials eliciting alignment (and 1/12 of all trials). For another group of participants, the non-native confederate produced no errors. In Experiment 2, Hypothesis 1 predicts more lexical alignment with the non-native than with the native confederate. Hypothesis 3 predicts that such an effect would be larger for the confederate who produced errors. Note that we expected participants never to align to the erroneous gender markers – just to the picture names themselves.

We acknowledge that there are reasons to treat with caution the use of confederates in dialogue experiments (Kuhlen and Brennan, 2013). We used confederates because dispreferred names are crucial for the detection of alignment but hard to elicit spontaneously. Kuhlen and Brennan note that in such a case the use of confederates may be the only feasible approach (p.56, p.65). In accordance with their recommendation, we document relevant details about confederate knowledge and behavior in the Method section, and discuss how confederate participation could have affected the results in the General Discussion.

We also note that the current experiments sought to determine the mechanisms of a specific type of language use and thus had features that made them quite different from many natural dialogues. Some relevant features were the overall ease of the task, the preponderance of referring expressions (over non-referential language) together with the otherwise limited conceptual content, the lack of grounding in most conditions, and the formulaic grounding in the Feedback group of

Experiment 1; we consider in the General Discussion how these features could have affected the studied mechanisms.

2. Experiment 1: Lexical alignment to native and non-native speakers of English with or without feedback of comprehension success

In this experiment, we contrasted native English speakers' lexical alignment to native and non-native confederates in a picture-matching task. Half of the participants and confederates gave and received feedback about successfully matching the pictures, while the other half matched pictures without such feedback.

2.1 Method

In this section, we provide details about the participants and confederates, the materials and procedure, and the design, coding, and data analysis in Experiment 1.

2.1.1 PARTICIPANTS

Forty-eight students from the University of Edinburgh community were paid to take part. All participants were native speakers of English. The data from one participant in the No-feedback group was excluded from the analyses because of a counterbalancing error. There were thus 24 participants in the Feedback group, and 23 in the No-feedback group.

2.1.2 CONFEDERATES

Four confederates who were non-native speakers of English were selected on the basis of a pretest. Nineteen non-native speakers and one native speaker of English were recorded producing the preferred and dispreferred names of the experimental items described below (36 words in total). After, twelve further participants from the University of Edinburgh community who did not participate in the main experiment listened to these recordings (in randomized order for each rater) and indicated the perceived strength of the speaker's foreign accent on a Likert scale (1 meant "extremely strong foreign accent" and 7 meant "native English accent"). Four non-native speakers with medium-strong accents (to ensure intelligibility; average ratings of 3.25, 3.33, 3.75 and 4.08) were selected for participation in the main experiment. *T*-tests on the ratings of each two pairs of speakers for all possible pairings indicated that the ratings for all four speakers did not differ [all $ps > .17$].

Four further native speakers of English were selected to act as the native confederates. Each of these speakers was paired with one of the non-native speakers to form same-gender pairs (three female and one male pairs). All confederates were paid for their time. The confederates were aware of the experimental hypotheses. Confederates arrived when their turn was scheduled and were not present in the room when another confederate was performing the task – thus, they did not have information about the behavior of the confederate they were paired with. Each confederate performed the task 11 or 12 times (see below).

2.1.3 MATERIALS AND PROCEDURE

The materials consisted of black-and-white line drawings and were the same as in Branigan et al. (2011). There were 18 experimental pictures, each of which could be named with both a preferred

and a dispreferred but acceptable name (for example, *lamp-light*). The names were normed by Branigan et al. (2011) for the same population as tested here. In addition, there were 154 filler pictures which only had one name and appeared between one and eight times in the combined participants' and confederate's naming and matching sets.

The materials were presented in binders, one for the participant and one for the confederate (see Figure 1 for the experimental set-up). In the participant binder, two pictures were displayed on each page (an A4 sheet inside a plastic envelope). On a naming turn, the picture to be named was indicated by a thicker black border, and appeared on the left side on half of the trials, and on the right side on the other half. On a matching turn, both pictures had the same thickness border. Matching was performed by placing a sticker on one of the two pictures on every matching sheet. In the confederate binder, each naming turn had a single word printed in the center of the page. Matching trials had the same structure as participants'.

On each naming or matching turn, an individual filler picture appeared next to an experimental picture. On critical turns, the confederate named an experimental picture with either its preferred or its dispreferred name (counterbalanced across items), and the participant located it on the corresponding matching page. After two filler turns (participant naming trial, confederate naming trial), the participant had to name the same experimental picture (again, presented together with a filler picture). There were 14 filler turns at the beginning of the experiment (the first four were presented as practice). There were two such beginning-filler sets, counterbalanced across lists, confederate orders and the four confederate pairs. After that, six filler turns separated each 4-trial sequence involving experimental pictures; those were kept constant within lists.

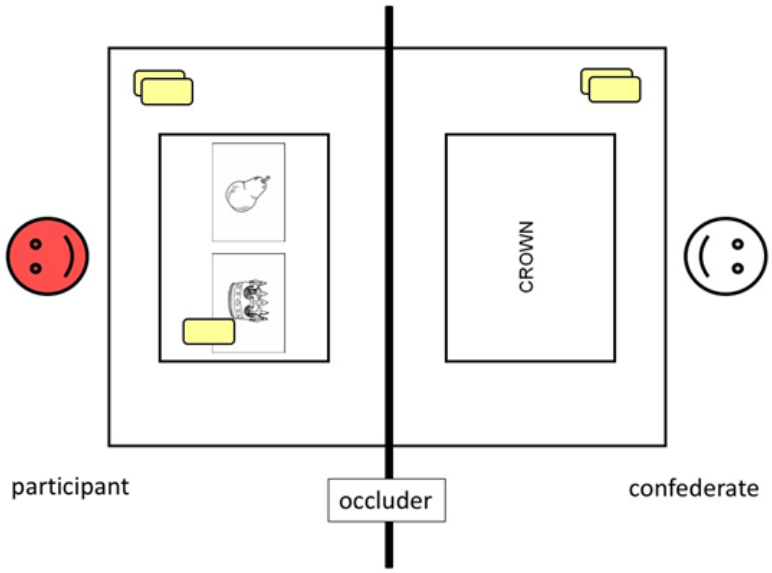

Figure 1: Experimental paradigm for the picture-matching game used in Experiment 1. The confederate has performed a naming turn, and the participant has indicated their choice using a sticker (indicated as a yellow rectangle).

There were eight item lists. Four lists (List set A) were created by counterbalancing confederate names for the experimental pictures (preferred, dispreferred) and the side on which experimental items appeared on participant naming and matching trials (left, right). The content and order of filler trials was the same across these four lists except the beginning filler sequence, as explained above). Four further lists (List set B) were created by randomizing the order of the filler trials and experimental sequences (but the order of items within an experimental sequence was the same as in the first four lists). Across the latter four lists, the content and order of filler trials was also kept the same. Within a single list, there were 152 confederate naming trials, and 152 participant naming trials.

Each participant performed the matching game with both confederates within a confederate pair, one a native speaker of English, and the other a non-native speaker, in counterbalanced order. Twelve participants performed the experiment with each of the first three confederate pairs, and 11 participants performed the experiment with the last confederate pair; confederate pairs were equally distributed across the Feedback and No-feedback groups. Each participant performed the experiment with a same-gender confederate pair. If the first part of the experiment (interaction with the first confederate) used a list from List set A, the second part of the experiment (interaction with the second confederate) used a list from List set B that contained the alternative names for experimental items (for example, if the first confederate named a picture of a basket with its dispreferred name *hamper*, the second confederate named the same picture *basket*). The order of list sets was counterbalanced across participants.

Upon arrival, participants and confederates were assigned to their seats, to ensure that confederates did not always head to the same side out of habit, which could indicate that they had been in the room before. Seat order was counterbalanced, such that half of the time confederates sat facing the window and half of the time they sat facing the door. Participants and confederates read written instructions that explained the procedure involving turn-taking with the other person to name and match pictures. Participants were not told that they would perform the task with two different partners at the beginning, but only upon arrival of the second confederate (approximately 30 minutes after the beginning of the experimental session, the time taken to complete the task with the first confederate). The instructions explained that to-be-named pictures were surrounded by a thicker black border and that matching was to be performed by placing a sticker on the matching picture. Additionally, participants and confederates in the Feedback group were instructed to give verbal feedback of which pictures they matched. They did this by saying "left" or "right", depending on the position of what they thought was the correct matching picture on a given matching sheet (and they were always correct because the task did not leave any possibility for confusion for a native speaker). Importantly, they were also instructed to pay attention to this verbal feedback when it came from the other person, and place a sticker on their own naming sheet to indicate which picture their partner had selected. This was done to encourage participants to attend to the feedback. Once participants or confederates performed a turn (for a naming turn, produced a name, heard the feedback from the other participant and placed a sticker accordingly; for a matching turn, heard the name produced by the other person, decided on the correct matching picture, placed a sticker on it and said "left" or "right" to indicate its position), both participants turned a page in their binders; this indicated the beginning of the next trial. The procedure for participants in the No-feedback group was identical, except that no feedback was given (or monitored) by either participants or confederates with respect to their matching choices. All experimental sessions were recorded for subsequent verification.

2.1.4 DESIGN, CODING, AND DATA ANALYSIS

Responses were transcribed in real time by the first author and were recorded for subsequent verification. We report here analyses of the responses following dispreferred names produced by confederates, to the extent that they carry condition differences in alignment effects (while preferred responses following preferred confederate names were close to ceiling, 95.3%). Analyses of responses following preferred confederate names are reported in Appendix A. The data were analyzed with logistic mixed-effects regression modeling. Dispreferred responses (matching the names produced by confederates) were coded as 1, and preferred responses (different from the names produced by confederates) were coded as 0. The fixed predictors were Confederate type (native, coded as 0.5, and non-native, coded as -0.5), Feedback group type (feedback, coded as 0.5, and no feedback, coded as -0.5), Confederate order group (native-first, coded as 0.5, and non-native first, coded as -0.5) and their interactions.

The Confederate order group factor was included in the analyses because of the following considerations. First, evidence from structural alignment suggests that speakers adapt to the frequency of alternatives in an experiment, resulting in a gradual decrease of alignment throughout an experimental session (Fine and Jaeger, 2013). In our experiment, confederate order was counterbalanced (half of the participants performed the picture-matching task with a native confederate first and a non-native confederate afterwards, and for half of the participants this assignment was reversed), but alignment decrease because of adaptation might be differentially influenced by confederate type. Further, in this experiment the second confederate produced the alternative name to the name produced by the first confederate. For example, if the native confederate was the first to interact with a participant and produced a preferred name (for example, *lamp*), the non-native confederate subsequently named the same picture with the dispreferred name (for example, *light*), or vice versa. Because of this design, participants might align more readily with a name (thus, accept a label for a given object, for example *lamp*) if they have not yet themselves produced a name for that object, but then be less likely to adopt a different name once they have named this object because speakers tend to repeat themselves (see Experiment 3 in Branigan et al., 2011; Wheeldon and Monsell, 1992; Brennan and Clark, 1996, though note that speakers in dialogue *can* change the way they refer to objects with each different conversation partner). For both of these reasons, it is possible that alignment to the native and non-native confederates was differentially impacted by the position of the respective interaction within our experiment. Note that the decision to include the Confederate order group as factor was post-hoc but analyses without it produced an equivalent pattern of results for the remaining factors.

The models were run using the *glmer* function in the lmerTest package (version 2.0-33, lme4 version 1.1-13) in R (version 3.4.1). To aid convergence, the "bobyqa" optimizer was used. All models initially had the maximal random-effects structure justified by the design (Barr et al., 2013). In case of non-convergence of the full model, model simplification was performed by first removing random-effects correlations and subsequently removing in a step-wise manner the random effects accounting for least variance (random slopes were removed before random intercepts). To shed light on significant interactions, further models were run on subsets of the data; these models are described in the Results section below.

The data are publicly available at `https://osf.io/7rtgd/`, and analyses scripts will be offered upon request.

2.2 Results and discussion

The by-participant percentage aligned responses are plotted in Figure 2 and the statistical models results are reported in Table 1. We first report the results of the main model described above. After, we report analyses breaking down the data into Feedback and No-feedback groups, and, within each of these, into Native-first and Non-native-first groups. These breakdowns were done because differential effects of feedback across confederate types is predicted under Hypothesis 2 (and visible numerically on Figure 2) but power may not be sufficient to detect the interactions of interest.

The main analysis revealed globally more alignment for dispreferred responses to the non-native (39.7%) than to the native confederate (33.6%; the main effect of Confederate type was significant).

Further, participants in the Native-first group aligned more overall (42.8%) than participants in the Non-native-first group (30.9%; the main effect of Confederate order group was significant). This difference, however, was due to participants in the No-feedback group (there was a significant interaction between Feedback group and Confederate order group). Specifically, Feedback-group participants who interacted with the native confederate first showed similar overall alignment to Feedback-group participants who interacted with the non-native confederate first (a difference of 1.7%), while No-feedback-group participants who interacted with the native confederate first showed 25% more alignment than No-feedback-group participants who interacted with the non-native confederate first. These differences might suggest an effect of confederate type whereby interacting with a non-native confederate first globally reduced the rate of alignment to dispreferred names throughout an experimental session. However, we believe these differences mostly likely stem from individual differences in alignment to dispreferred names because they come from a between-participant comparison; as such, we do not interpret them further.

Participants also aligned more to the confederate with whom they interacted first, regardless of confederate type (there was a significant interaction between Confederate type and Confederate order group). But this tendency was more pronounced when the first conversation partner was the non-native confederate (17% more alignment with the non-native than with the native confederate) than when it was the native confederate (5.3% more alignment with the native than with the non-native confederate; note that this is a between-participants comparison).

We then ran separate models on the data of each feedback group type, with the fixed predictors Confederate type, Confederate order and their interaction. Participants in the Feedback group aligned overall to a similar extent to the native and non-native confederates (no significant main effect of Confederate type). However, they had a tendency to align more with the confederate they interacted with first (a significant interaction between Confederate Type and Confederate order group). In other words, when there was no need for concern about comprehension success, there was less alignment with the second consecutive conversation partner than with the first, independent of that partner's native language. (This tendency was stronger for the Non-native-first group, who showed 17.3% more alignment with the first non-native confederate, and a significant simple effect of Confederate type, than for the Native-first group, who showed 13.1% more alignment with the first native confederate but no significant simple effect of Confederate type.) Importantly, participants in the No-feedback condition aligned more to the non-native than to the native confederate (a significant effect of Confederate type). Further, as indicated by the main analysis, participants in the Native-first group aligned more overall than participants in the Non-native-first group (a significant effect of Confederate order group).

Simple effects models on the data for the two Confederate order groups with Confederate type as a fixed predictor showed more alignment with the non-native than with the native confederate (a difference of 16.7%) for participants who interacted first with the non-native confederate (a significant effect of Confederate type), but no difference in alignment to the two confederates (1.9%) for participants who interacted first with the native confederate (no significant effect of Confederate type). These analyses suggest that, without evidence for comprehension success, the tendency to align less to dispreferred names from a second partner occurred only when this partner was the native confederate; when the non-native confederate was second, alignment did not differ from alignment with the first confederate.

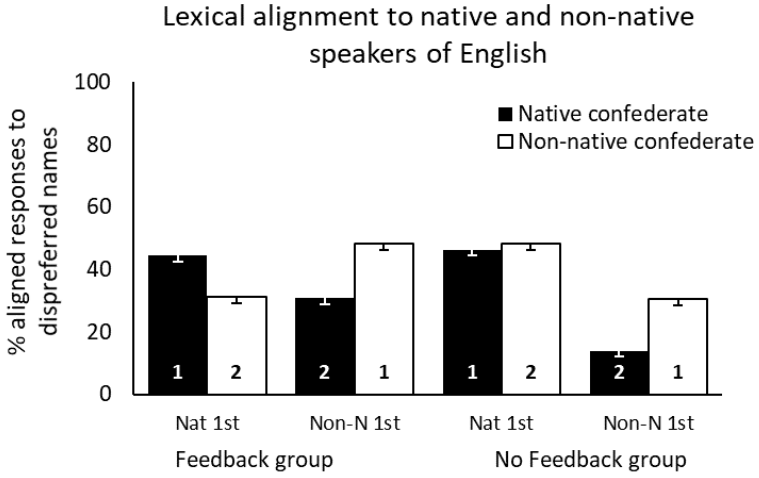

Figure 2: Mean by-participant percentage alignment effects by feedback group type, confederate order group and confederate type in Experiment 1. Error bars represent standard error (computed from by-participant means). Nat 1st: Native-first group; Non-N 1st: Non-native-first group. The numbers inside the bars indicate condition order in the experiment.

In sum, Experiment 1 revealed that, with feedback, participants aligned less to their second than to their first partner, regardless of their native language. Without feedback, participants aligned less to their second than to their first partner when the second partner was a native speaker, but maintained alignment at its initial level when the second partner was a non-native speaker. These results support Hypothesis 1, which is about the direction of the confederate type effect: Under certain conditions, alignment to the non-native speaker was larger (not smaller) than that of the native speaker. This directionality supports an influence of audience design in alignment to non-native speakers. The results also support Hypothesis 2 in that such effects appeared only in the absence of feedback. This pattern is also consistent with an influence of audience design - but suggests that audience design is used only when deemed necessary (i.e., in the absence of evidence for comprehension success).

Additionally, we found that alignment tended to decrease from the first to the second consecutive conversation partner. This effect could have sources in participants' adaptation to the frequency of dispreferred alternatives within the experiment (for such an inverse preference effect in structural alignment, see Jaeger and Snider, 2013), as well as participants' tendency to self-repeat (Branigan

et al., 2011; Wheeldon and Monsell, 1992): If participants aligned with the first confederate (for example, said *lamp*), they might have then been less willing to switch to a different name with the second confederate (for example, say *light*).

Model	Predictors	*Estimate*	*SE*	*z*	*p*
Main model	Confederate type	-.41	.17	-2.33	.02
	Feedback group	.29	.39	.75	.45
	Confederate order group	.77	.39	1.99	.047
	Confederate type * Feedback group	.51	.35	1.45	.15
	Confederate type * Confederate order group	1.40	.42	3.37	<.001
	Feedback group * Confederate order group	-1.67	.77	-2.15	.03
	Confederate type * Feedback group * Confederate order	.31	.72	.44	.66
Feedback group	Confederate type	-.14	.29	-.50	.62
	Confederate order group	-.07	.60	-.11	.91
	Confederate type * Confederate order group	1.58	.62	2.55	.01
Native-first group	Confederate type	1.54	1.08	1.43	.15
Non-native-first group	Confederate type	-.93	.45	-2.05	.04
No-feedback group	Confederate type	-.69	.32	-2.19	.03
	Confederate order group	1.62	.53	3.06	.002
	Confederate type * Confederate order group	1.22	.69	1.76	.08
Native-first group	Confederate type	-.07	.47	-.14	.89
Non-native-first group	Confederate type	-1.12	.37	-3.04	.002

Table 1: Results of the LMER main model of dispreferred responses in Experiment 1. *Note: Grey rows indicate significant effects.*

3. Experiment 2: Lexical alignment to native and non-native speakers of Spanish with or without grammatical errors

In this experiment, we contrasted native Spanish speakers' lexical alignment to dispreferred names produced by native and non-native confederates in a different version of the picture-matching game. The non-native confederate made grammatical (gender) errors on half of the target pictures with one participant group and made no errors with another group, so that we could see if errors from the conversation partner (a signal of lower communicative competence) would increase alignment. An additional aim of this experiment was to determine the extent to which the findings of Experiment 1 would generalize to another population and language.

3.1 Method

In this section, we provide details about the participants and confederates, the materials and procedure, and the design, coding, and data analysis in Experiment 2.

3.1.1 PARTICIPANTS

Sixty-four undergraduate students from the University of Barcelona, (Barcelona, Spain), took part in exchange for course credit. All participants were native speakers of Castilian Spanish (spoken in their home by both parents, and currently spoken by participants more than 50% of the time). Participants also spoke Catalan (the population of Barcelona is almost entirely bilingual). For forty participants, the non-native confederate made no grammatical errors (No-error group), and for twenty-four participants the confederate produced grammatical errors on half of the experimental items (Error group). We acknowledge the uneven number of participants in the two error groups. Participant testing in the current No-error group was originally conducted as two separate experiments, which we now combine for greater power. The two combined experiments followed an identical procedure. The only differences were that the confederates were different individuals, and that two items, described below, were changed. Sample sizes were originally chosen to approximate those in the study of Branigan et al. (2011), which were 16 in Experiments 1 and 3, 20 in Experiment 5, 24 in Experiment 4 and 32 in Experiment 2. The uneven number of participants in the two groups was adjusted in the statistical analyses by centering the numerically coded Error group factor around the mean.

3.1.2 CONFEDERATES

The confederates for the Error group and the first 16 participants in the No-error group were two female graduate students from the University of Barcelona, a native speaker of Castilian Spanish and a native speaker of German. The confederates for the remaining participants in the No-error group were a female graduate student from the University of Barcelona, a native speaker of Castilian Spanish, and a female student from the University of Barcelona community, a native speaker of American English. The confederates were aware of the experimental hypotheses. They were not present in the room when another confederate was performing the task. Each confederate performed the task 24 or 40 times. Confederates saw pictures with the names written underneath for both experimental and filler items.

3.1.3 MATERIALS AND PROCEDURE

The experimental items were 22 further black-and-white line drawings, which could be named with a preferred and a dispreferred name (for example, *puño* [Sp. fist] – *mano* [Sp. hand]). Both names for all stimuli are listed in Appendix B, but note that confederates produced only dispreferred names in this experiment. The dispreferred names were always more frequent than the preferred names, to lend credibility to the fact that they were spontaneously produced by non-native speakers (likely not expected to know low-frequency words). The experimental pictures were selected on the basis of a pretest from an initial set of 53 pictures, as follows.

Thirty-two further participants from the same population and who did not take part in the main experiment were asked to name the 53 pictures. They also rated the appropriateness of the dispreferred name for each picture on a 10-point scale. These tasks were presented in counterbalanced order. A stimulus set of 24 of these items was then tested with seven pilot participants using the paradigm in Experiment 1. These 24 stimuli involved two changes from the normed ones: the picture of a pigeon was replaced with a picture of a parrot (both had *pájaro* [Sp. bird] as dispreferred name); and the dispreferred name *via* [Sp. way] for the preferred *carretera* [Sp. freeway] was replaced with *camino* [Sp. road]. The pilot revealed numerically greater alignment to the non-native

than to the native confederate. The paradigm was changed to the one described below to enhance ecological validity and reinforce beliefs that the task would be hard for a non-native speaker.

From the items in the pilot, four were removed and three were replaced, for example because some of the pictures were hard to recognize (e.g., tree trunk; braid), leaving 20 items. This stimulus set was used with 16 participants in the No-error group. The preferred names in this set were spontaneously produced 96.4% of the time on average (SD = 4.8%; norming data); the corresponding dispreferred names had medium acceptability (M = 6.11, SD = 1.04). For the remaining 24 participants in the No-error group and all participants in the Error group, two further pictures (baby and bride) were replaced (with fridge and earring) because gender errors on entities with a natural gender (for example, *una niño* [a (fem.) male child] or *un mujer* [a (masc.) woman] would have appeared exceptionally strange.

The experimental set-up (see Figure 3) consisted of two description sets (two sets of pictures (A) arranged on three-by-four-grids on ten PowerPoint slides per set and then printed on A4 sheets of paper), two match sets (two sets of loose pictures (B), prepared by cutting up different copies of the description sets into individual pictures), and two sets of blank templates (C) printed on A4 sheets of paper with numbered three-by-four-grids on each of them. Each of the ten sheets in one description set, together with their corresponding loose pictures, made one round of the game for one participant. The confederate's description sheet additionally contained all picture names written under their corresponding pictures.

To create the description sets, 20 experimental pictures, together with 176 filler pictures, were arranged on ten three-by-four grids drawn on Microsoft Powerpoint slides (some fillers occurred on both the participant and confederate description sets, and some were unique to either the participant or confederate set). There were two experimental pictures on each slide in each round of the game. An experimental picture always occurred first in the confederate's description sheet and was preceded by at least two filler pictures. The same picture then occurred on the participant's description sheet, after two filler naming turns (participant's and confederate's; this separation was necessary to make the repetition less obvious). Each participant saw and named the 20 experimental items only once with each confederate. For this purpose, the 10 rounds were divided into two sets of five rounds, one to be performed with each confederate.

The filler pictures in both the confederate's and participant's description sets contained pictures whose names were similar in meaning or form to the target names (for example, toothpaste, toothbrush, dentist, for the target *denture* (preferred) – *teeth* (dispreferred)), to make the matching task more challenging, presumably especially for a non-native speaker. The pictures in the confederate's description set had more frequent names than the ones in the participant's description set, to reinforce the idea that the non-native confederate had limited vocabulary (that is, was not fully proficient in Spanish); but note that these pictures were the same for the native confederate.

On every round, participant and confederate each received a description sheet, twelve match pictures from their match set (containing the pictures from the other person's current description set) and a blank template. They then took turns to name a picture from their description set (moving horizontally from left to right), and then to find the picture just named by the other person and place it on its corresponding position on the blank template. The confederate and participant each went first on half of the rounds, such that the participant went first on all rounds performed with the first confederate and second on all rounds performed with the second confederate, or vice versa.

To make the task somewhat more naturalistic, participants and confederates were instructed to produce a whole sentence instead of just the picture name. An example exchange went as follows:

Confederate: *Mi primer dibujo es un libro* [My first picture is a book]
Participant (after matching): *Mi primer dibujo es una pelota* [My first picture is a ball]
Confederate (after matching): *Mi número dos es un espejo* [My number two is a mirror]
Participant (after matching): *Mi número dos es una vela* [My number two is a candle]

On critical naming sequences (independently of who had the first turn in the round), the confederate named an experimental picture, always using the dispreferred name (for example, *mesa* [Sp. table] for a picture of a desk). This was followed by two filler turns (for example, the participant then named a picture of a chair, and the confederate named a picture of a bookcase), after which it was the participant's turn to name the experimental picture (desk in this example). The experimenter, seated behind the participant and confederate, noted the name produced by the participant.

Additionally, for the Error group, the non-native confederate made a grammatical error on one experimental trial per round (thus, on half of the experimental trials in total). The errors consisted in producing feminine indefinite determiners for words of masculine gender (for example, **un terraza* [a (masc.) terrace]; **un mano* [a (masc.) hand]) or vice versa (for example, **una pollo* [a (fem.) chicken; **una juguete* [a (fem.) toy].

There were eight experimental versions for the No-error group, obtained by crossing confederate order (native first or non-native first), first-turn assignment (confederate or participant), and order of experimental items within a round. For example, if the picture of a highway was the first experimental item and the picture of an ambulance was the second experimental item on a given round for half of the participants, the ambulance was first and the highway second for the other half. There were additional eight versions for the Error group, obtained by crossing the presence or absence of an error on each experimental item. Further, half of the participants in both groups saw one 5-round set first and half saw the other 5-round set first, but this factor was not fully counterbalanced (that is, no additional groups were created beyond the ones described above). The order of the rounds within each 5-round set was randomly varied among participants.

Upon arrival in the lab, the participant and first confederate were seated side by side at a table with an occluder between them, such that they were not able to see each other's picture sets (see Figure 3). As in Experiment 1, the second confederate arrived later, when the task with the first confederate was completed, and participants were not told in advance that they would interact with two confederates. They read written instructions which informed them that the experiment involved a collaborative game and then stepped them through the procedure. The instructions further stated that each participant would receive "something sweet" (a KitKat bar) if both of them completed all rounds correctly, to further encourage a collaborative spirit and incentivize them to pay attention to their and their partner's performance. The game with each confederate began with a practice round consisting of six description pictures. This was followed by five experimental rounds, each consisting of 12 pictures. Once a round was completed, the experimenter verified that it was completed correctly (performance was at ceiling). The experimental sessions were recorded for subsequent verification.

3.1.4 DESIGN, CODING, AND DATA ANALYSIS

Responses were coded as in Experiment 1. The dependent variable was aligned responses, which were all dispreferred responses. The fixed predictors in the main model were Confederate type

(native, coded as 0.5, and non-native, coded as -0.5), Confederate error group (no-error group, coded as -0.5, and error group, coded as 0.5), Confederate order group (native-first, coded as 0.5, and non-native first, coded as -0.5) and their interactions. Confederate order group (a predictor determined post-hoc) was included in the model for comparison with Experiment 1; a model without this factor and its interactions produced an identical pattern of results for the other factors. To shed light on interaction terms, further models (described below) were run on subsets of the data.

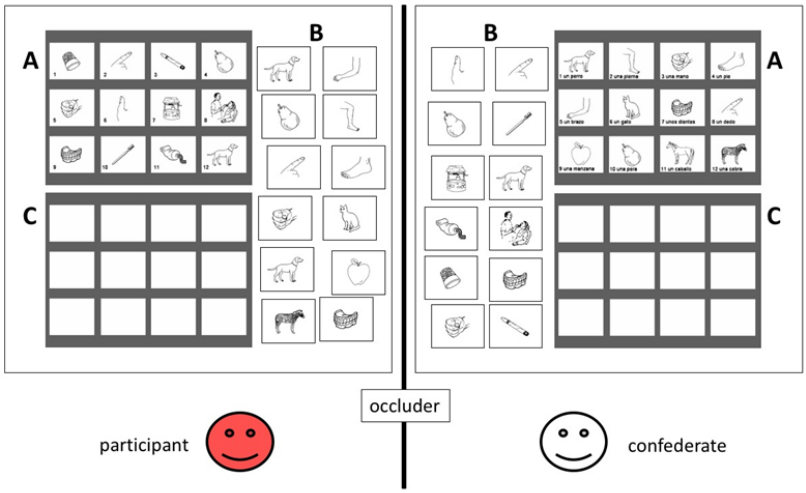

Figure 3: Experimental paradigm for the picture-matching game in Experiment 2.

3.2 Results and discussion

The by-participant mean proportions of dispreferred responses by confederate type, error group and confederate order group are plotted in Figure 4, and the statistical models results are reported in Table 2.

The main analysis revealed that participants produced more dispreferred responses overall when they interacted with the non-native confederate (40%) than when they interacted with the native confederate (31.1%) – that is, there was more alignment with the non-native than with the native confederate (a significant main effect of Confederate Type).

Further, participants experiencing error-free descriptions from the non-native confederate (No-error group) showed more overall alignment (41.3%) than participants experiencing gender errors from the non-native confederate (Error group; 26.1%; Confederate error group was a significant predictor). However, when the native confederate was first, participants in the No-error group aligned more even to the native confederate relative to participants in the Error group in the same condition (*Estimate* = -.79, *SE* = .40, z = -1.98, p = .048); the linguistic behaviour of the native confederate did not differ between the two Confederate error groups. This suggests that the alignment differences between the two groups were most likely due to random variability.

We further report analyses of subsets of the data parallel to those we conducted in Experiment 1. First, we conducted separate analyses of each confederate error group, with Confederate type, Confederate order group and their interaction as fixed predictors. For the Error group, this analysis showed that alignment with the non-native (23.0%) and native confederates (29.2%) did not differ (no main effect of Confederate type). However, for the No-error group, there was significantly more alignment with the non-native (46.5%) than with the native confederate (36.0%) (a main effect of Confederate type). The simple effects of Confederate order group for the No-error group further showed more alignment to the non-native than to the native confederate for participants who first interacted with the non-native confederate, but similar alignment to the non-native and native confederates for participants who first interacted with the native confederate. There were no other significant effects.

In sum, participants aligned their lexical choice more with the non-native than with the native confederate when the non-native confederate did not produce any grammatical errors. These results support Hypothesis 1 and point to a role of audience design in lexical alignment to non-native speakers in the structured task-based dialogue studied here. They also replicate with a different population and in a different language the greater alignment to non-native than to native speakers without feedback found in Experiment 1.

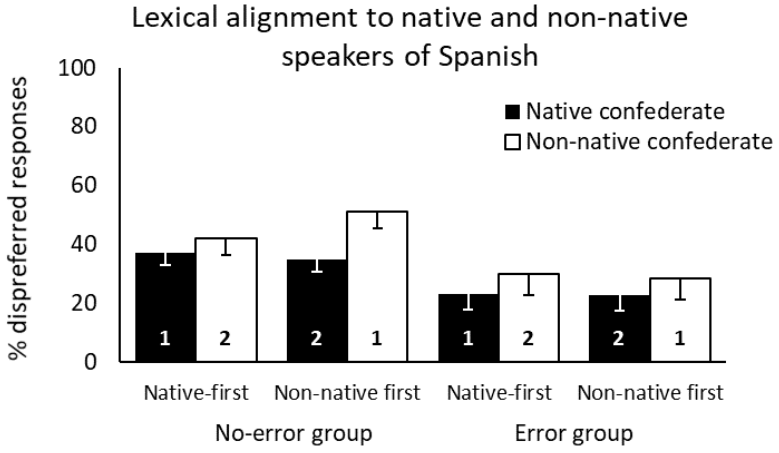

Figure 4: Mean by-participant percentage aligned (dispreferred) responses by error group, confederate order group and confederate type in Experiment 2. Error bars represent standard error (computed from by-participant means). The numbers inside the bars indicate condition order in the experiment.

However, Experiment 2 did not support Hypothesis 3: When the non-native confederate produced grammatical errors, alignment was similar to alignment with the native confederate (instead of being even larger than when the non-native confederate did not make errors). This finding suggests that the presence of errors does not uniquely signal the need for enhanced lexical alignment to ensure comprehension success. Instead, we suggest it may introduce competing motives, such as the desire to serve as an example of correct language use (i.e., model correct behavior).

We probed further into whether such behavior affected only trials with errors or the task-based exchange globally, and we found support for the latter. That is, we compared alignment to the non-native confederate on trials with errors (29.8%) to that on trials without errors (31.2%) and did not find a statistical difference between the two (*Estimate* = -.30, *SE* = .31, z = -.97, p = .33; this model had only a random intercept for items). We discuss possible roles of grammatical errors for lexical alignment to non-native speakers in structured task-based dialogue in the General Discussion.

Lastly, Experiment 2 did not reveal the second-confederate alignment reduction found in Experiment 1. One plausible explanation for this pattern is that participants in Experiment 2 were exposed to and produced names for the experimental pictures with only one confederate (that is, never had to name the same picture twice). We note, however, that there were many other differences between the two experiments.

Model	Predictors	*Estimate*	*SE*	z	p
Main model	Confederate type	-.51	.14	-3.75	<.001
	Confederate error group	-.85	.25	-3.46	<.001
	Confederate order group	-.08	.24	-.36	.72
	Confederate type * Confederate error group	.19	.29	.67	.50
	Confederate type * Confederate order group	.34	.27	1.25	.21
	Confederate error group * Confederate order group	.33	.49	.67	.51
	Confederate type * Confederate error group * Confederate order group	-.67	.57	-1.16	.25
No-error group	Confederate type	-.63	.26	-2.38	.02
	Confederate order	-.23	.33	-.69	.49
	Confederate type * Confederate order	.65	.47	1.39	.17
Native-first group	Confederate type	-.29	.30	-.98	.33
Non-native-first group	Confederate type	-.92	.38	-2.41	.02
Error group	Confederate type	-.40	.34	-1.19	.24
	Confederate order	.07	.38	.19	.85
	Confederate type * Confederate order	-.05	.70	-.07	.94
Native-first group	Confederate type	-.45	.75	-.60	.55
Non-native-first group	Confederate type	-.11	.45	-.25	.80

Table 2: LMER results for Experiment 2

4. General Discussion

This study tested three hypotheses about the role of audience design in lexical-referential alignment with non-native speakers in two structured task-based dialogue experiments. In Experiment 1 with speakers of English, there was greater alignment to dispreferred names produced by non-native confederates than to such names produced by native confederates in a participant group that did not receive feedback about the confederates' comprehension success. However, there was equivalent alignment in another group that did receive such feedback. Experiment 2 with speakers of Spanish replicated the greater alignment to non-native confederates in the absence of feedback (shown only in analyses of subsets of the data), but found equivalent alignment for another group that received descriptions from the non-native speakers that included some grammatical errors. The native/non-native differences in both experiments were driven by maintenance of alignment at the initial rate when the second partner was a non-native speaker, relative to a reduction of alignment when the second partner was a native speaker.

These results support our hypothesis that, under certain conditions, alignment with the non-native speaker would be larger (and never smaller) than with the native speaker (Hypothesis 1). Such differences suggest an influence of a form of audience design (the desire to aid communicative success) on lexical-referential alignment with non-native speakers in structured task-based contexts, across speakers of different languages. Our findings are thus in line with studies suggesting that speakers can override their own production preferences to adapt their lexical choices to their partners', and that they tend to do so when they believe or see that their partners are less competent speakers or have less expertise in the relevant domain (Branigan et al., 2011; Isaacs and Clark, 1987). The finding that alignment with the non-native speaker was greater only when evidence about their performance was absent also supports our hypothesis that enhanced alignment to non-native speakers would be more, or only, present in conditions lacking grounding (Hypothesis 2). This result shows that audience-design driven alignment is not indiscriminate, but adapts to the specifics of the situation (i.e., uncertainty about one's conversation partner's comprehension success).

Without feedback (Experiment 1) and without confederate errors (Experiment 2), alignment decreased from the initial levels when the native confederate was second, but remained unchanged when the non-native confederate was second. This finding raises the possibility that participants initially intentionally chose to align at a high rate before they were able to assess the overall difficulty of the task (that is, at the beginning of the experiment), but by the second block had realized that the task was easy enough for a native speaker to perform and there was no need for extra care in their referential choice (that is, alignment). However, when introduced to the non-native confederate in the second block and in the absence of evidence of their comprehension success, they might have wanted to ensure that the non-native confederate was able to perform the task, and hence continued to align at the initial level. This possibility further suggests a role for audience design, in that it attributes adjustments to the magnitude of alignment to top-down decision-making processes.

These findings also fit with broader evidence that speakers engage in partner modeling, or adaptation of one's own language production based on expectations of the addressee's knowledge (Horton and Gerrig, 2002). Here, we show its influence on alignment behavior even though evidence suggests that it is a resource-demanding process (Horton and Keysar, 1996; Vogels et al., 2015).

Interestingly, the results of Experiment 2 did not support our hypothesis that the greater alignment with the non-native confederate would increase still further when this confederate gave descriptions with grammatical errors relative to when they gave descriptions without errors (Hypothesis 3). Such an increase might have been expected under an audience design account because grammatical errors indicate even more limited competence in the target language. There could be several possible explanations for why the predicted increase was not found.

First, the presence of grammatical errors may have triggered a desire in participants to correct non-native speakers' errors or infelicities and thus model correct usage of the language (see e.g., Kurhila, 2001). Consistent with this, participants never produced incorrect gender markers themselves, even when they reused the dispreferred picture names on trials with errors (to the recollection of the experimenter, the first author; the recordings of the experiments are no longer available to verify). The act of always producing correct gender markers, thus correcting the confederate's gender errors, may have suggested or intensified the decision to also suggest better names for the pictures. We note that such behaviour is ultimately collaborative (and likely kept in check by the fear of appearing rude or excessively pedantic), in that it aims to improve non-native speakers' ability to function in the target language.

Another possibility (less likely in the current setting because it was not explicitly manipulated) is that native speakers did not want to imitate the behaviour of speakers who made language errors. Divergence from the speech of one's conversation partner can be used to indicate disaffiliation (Bourhis et al., 1979; Doise et al., 1976; Ludlow, 2014). This possibility is consistent with results from studies on structural alignment. For example, Heyselaar et al. (2017) showed that participants structurally aligned less with avatars that had a computerized voice and did not exhibit typical interactional human behavior such as facial expressions and looks to the conversation partner, relative to human-like avatars. Weatherholtz et al. (2014) reported that the greater the perceived distance between recorded speakers' accents and participants' own, the less participants structurally aligned with this speaker.

We caution, however, that findings of studies of one type of alignment may not generalize to studies of a different type of alignment, because different levels of linguistic representation might be differentially sensitive to different mechanisms driving alignment. For example, syntactic misalignment would not normally result in miscomprehension, because a speaker can assume that even a less proficient conversation partner has some mastery of common syntactic alternations, as well as because utterance meaning is at least partially conveyed in lexical items and is not only carried by syntactic structure. In contrast, naming an object in a way the conversation partner does not understand can result in direct lack of comprehension. Further research is thus needed to systematically examine the interplay of mechanisms driving alignment at the different linguistic levels, and their similarities and differences.

Our results have implications for the increasingly common conversations between native and non-native speakers in real life. First, they suggest that the communicative difficulties faced by non-native speakers might be at least partially offset by native speakers aligning to their lexical choices. Second, the alignment of native speakers to non-native speech might, in the long run, have implications for language change (especially of English, which is spoken by many non-native speakers around the world): After aligning to certain non-native uses of words or expressions, native speakers might become more likely to adopt them even in conversations with native speakers, who might in turn align to them and adopt them in their own vocabularies. To gain more insight into this process, it would be necessary to investigate whether those native speakers' lexical choices that result from alignment with non-native speakers tend to persist with other speakers or remain partner-specific.

We note that the structured, task-based dialogue in our experiments had many differences from natural dialogue. One such difference was the participation of confederates. Kuhlen and Brennan (2013) point out that confederates may unwittingly behave in a way that biases participants to engage in the hypothesized behaviors, that they may show more knowledge than is assumed for somebody unfamiliar with the task, that participants may behave differently if they suspect that they are speaking to lab assistants rather than naïve participants like themselves, and that confederates that have performed the experiment multiple times and are reading from a script do not behave naturally. Features of our design undermine some of these concerns but we cannot completely exclude them.

First, the critical comparison was across two different confederates rather than between two conditions involving the same confederate engaging in different behaviors. Second, the likelihood of confederate behavior that consistently biased in a single direction seems undermined by the fact that there were four confederate pairs in Experiment 1 and two confederate pairs in Experiment 2, and by the different patterns for the Feedback and No Feedback vs. Error and No-error groups (although we

cannot exclude the possibility that confederate behavior affected differences between the Feedback and No Feedback conditions). Third, confederates did not know how the other confederate in their (native/non-native) pair behaved because they were not in the room during the other confederate's part of the session. Fourth, if participants had suspected confederate status, they would likely have inferred that the non-native confederates did not need help to perform the task; as such, we should not have observed more alignment to the non-native confederates.

Apart from the participation of confederates, the referential communication task we used had a greater proportion of referential expressions than a typical conversation, but no other conceptual content. Lexical-referential alignment may thus be overall much less in a natural conversation because of the fewer opportunities for it to occur; consistent with this, more alignment was found in a corpus of task-oriented conversations than in a corpus of unconstrained telephone conversations (Reitter et al., 2006). As such, differences between alignment with native and non-native speakers may be obscured or reduced in such unconstrained contexts compared with our study.

Another relevant difference between many natural conversations and our study is the presence of overt grounding (which occurred in our study only in the Feedback group) or explicit clarification requests prompting overt explanations. As we have shown, overt evidence for comprehension success modulates alignment to non-native conversation partners, and may make alignment less necessary as a comprehension-ensuring strategy in more naturalistic situations.

Further, the social desirability of alignment (cf. Communiucation Accommodation Theory, Giles and Powesland, 1975; Hopkins and Branigan, 2020) was not specifically manipulated here but may be a strong determinant of lexical-referential alignment in real-life situations, many of which also have a greater emotional component than our experiments. As such, factors that exercise a stronger influence than audience design may ultimately determine alignment to non-native (and also native) speakers in many contexts.

Lastly, natural conversation may be substantially more cognitively effortful than the simple tasks used here – for example, because of the need to plan more complex utterances or concurrent activities. Too great a cognitive load or time pressure may reduce or eliminate audience design (for example, Horton and Keysar, 1996) and leave only alignment based on automatic priming. This may eliminate or reverse the greater alignment to non-native speakers found in our experiments. Considering these differences, our study shows that audience design is a mechanism that can act to enhance lexical-referential alignment to non-native relative to native speakers – but it remains to establish the extent of its influence in natural conversations.

We have concluded that audience design played a role in alignment to non-native speakers - but did it play a role in alignment to *native* speakers? We speculate that such alignment may involve audience design in at least some contexts, but that the underlying influence of automatic priming may be stronger for native than for non-native conversation partners. If so, we can tentatively hypothesize that alignment to non-native speakers is more cognitively demanding than alignment to native speakers. The greater cognitive demand would be imposed by the need to assess non-native speakers' knowledge and conversational needs, which may not be undertaken to such an extent with native speakers. This then predicts that a manipulation of cognitive load would show a greater reduction of alignment under load with non-native than with native speakers. This prediction, however, remains to be tested.

In sum, we compared lexical-referential alignment with non-native speakers to that with native speakers, to test three hypotheses about the role of audience design in alignment to non-native speakers. Our results from two different language populations suggest that, within the context

of our experiments, audience design led to greater alignment with non-native speakers, but only when communicative success was uncertain (Experiment 1). Grammatical errors produced by the non-native confederate did not increase alignment even further, despite evidencing lower communicative competence in the target language (Experiment 2). This result is instead consistent with an alternative collaborative strategy, the desire to demonstrate correct language use (specifically to non-native speakers). An implication of these results – although one that remains to be tested in more naturalistic settings – is that non-native speakers' production in conversation is ultimately facilitated by their native conversation partners, who act to ensure communicative success.

Acknolwedgements

Albert Costa died on 10th December, 2018. We dedicate this article to his memory. This research was supported by Spanish Government funds (grants SEJ05 62542CV00568007 and PSI 2008-01191/PSIC, awarded to Albert Costa, FPU fellowship AP2005-4496, awarded to Iva Ivanova) and an ESRC grant RES-062-23-0376, awarded to Holly Branigan and Martin Pickering. Heartfelt thanks go to Katy Bellamy, Anna Leonard Cook, Sarah 'Sez' Gordon, Wan-Yu Hung, George Kountouriotis, Nien Chen Lee, Oliver Stewart and Anna Vasilyeva for acting as confederates in Experiment 1, to Yolanda Garcia, Jennifer Klimowicz, Sara Rodriguez and Jasmin Sadat for acting as confederates in Experiment 2, and to Kyle Wolff for formatting the final version of the manuscript. We also thank the Editor and two anonymous reviewers for the many useful comments. The authors declare no conflict of interest.

Appendix A: Analysis of Responses to Preferred Confederate Names in Experiment 1

Analyses of responses to preferred confederate names were the same as the analyses of responses to dispreferred confederate names reported in the main text. The main logistic mixed-effects model had Confederate type (native, coded as 0.5, and non-native, coded as -0.5), Feedback group (feedback, coded as 0.5, and no feedback, coded as -0.5), Confederate order group (native-first, coded as 0.5, and non-native first, coded as -0.5) and their interactions as fixed predictors. Dispreferred responses were coded as 1, and preferred responses as 0. Further separate models on the data of each feedback group had Confederate type, Confederate order and their interaction as fixed predictors.

The results of the statistical models used to analyze these effects are reported in Table 3. Participants aligned to a similar extent with the native and non-native confederate (in the main model, the effect of Confederate type was not significant). Further, participants aligned more to the confederate with whom they interacted first, regardless of confederate type (there was a significant interaction between Confederate type and Confederate order group). Specifically, the Native-first group aligned more to the native than to the non-native confederate (a difference of 4.3%), while the Non-native-first group aligned less to the native than to the non-native confederate (a difference of 3.2%). These differences were further modulated by the presence of feedback (there was a significant three-way interaction between Confederate type, Feedback group and Confederate order group).

The separate models on the data of each feedback group indicated that the tendency to align more with the first confederate was carried by participants in the Feedback group (a significant interaction between Confederate Type and Confederate order group for participants in this group). Further separate models on the data of each Confederate order group suggested that this tendency was more robust for the Native-first group (a difference of 10.1% and a significant effect of Confederate order) than for the Non-native-first group (a difference of 4.6% but no significant effect of Confederate order.) For the No-feedback group, alignment to the two confederates was similar between the Native-first group (a difference of 0.9%) and the Non-native-first group (a difference of 1.9%; there were no significant effects).

Taken together, these results suggest that alignment with two consecutive conversation partners is influenced by interaction order, but only when there is no need for concern about communicative success. The tendency to align less with the second conversation partner in the presence of feedback was also attested in the analyses of dispreferred responses reported above. This tendency was possibly driven by increasing exposure to both alternatives throughout the experiment causing more varied responses (i.e., more noise in the system), as well as by participants' reluctance to switch to a different name once they had produced a name for a given object – even when it was the preferred name. But this tendency was unaffected by confederate native language only when participants received feedback of comprehension success. In line with the findings reported above, when comprehension success was unclear, this tendency disappeared, and alignment to the non-native confederates when they were the second conversation partners was 7.5% more than in the same situation in the Feedback condition (although note that this is a between-participant comparison). Considerations about comprehension success in the absence of feedback may have made alignment to preferred names more likely in general, even with native confederates.

Model	Predictors	*Estimate*	*SE*	*z*	*p*
Main model	Confederate type	.18	.40	.45	.66
	Feedback type group	-.03	.40	-.07	.95
	Confederate order group	.07	.40	.17	.86
	Confederate type * Feedback type group	1.12	.81	1.39	.16
	Confederate type * Confederate order group	1.91	.81	2.37	.02
	Feedback type group * Confederate order group	.44	.81	.54	.59
	Confederate type * Feedback type group * Confederate order group	3.23	1.62	2.00	.046
Feedback group	Confederate type	.78	.64	1.22	.22
	Confederate order group	.28	.64	.44	.66
	Confederate type * Confederate order group	3.36	1.29	2.61	.009
Native-first group	Confederate type	2.51	1.06	2.37	.02
Non-native-first group	Confederate type	-1.74	1.83	-.95	.34
No-feedback group	Confederate type	-.40	.52	-.77	.44
	Confederate order group	-.16	.52	-.31	.76
	Confederate type * Confederate order group	.58	1.20	.48	.63
Native-first group	Confederate type	-.22	.90	-.24	.81
Non-native-first group	Confederate type	-3.29	5.59	-.59	.56

Table 3: LMER analyses of responses to preferred names in Experiment 1

Appendix B: Experimental Items in Experiment 2

carretera [freeway] – camino [road]
ambulancia [ambulance] – coche [car]
copa [wine glass] – vaso [glass]
hamurguesa [hamburger] – bocadillo [sandwhich]
escritorio [desk] – mesa [table]
litera [bunk bed] – cama [bed]
imperdible [safety pin] – aguja [pin]
caramelo [candy] – dulce [sweets]
palmera [palm tree] – árbol [tree]
rosa [rose] – flor [flower]
balcón [balcony] – terraza [terrace]
cuadro [painting] – dibujo [drawing]
puño [fist] – mano [hand]
dentadura [denture] – dientes [teeth]
mochila [backpck] – bolsa [bag]
muñeca [doll] – juguete [toy]
gallina [hen] – pollo [chicken]
loro [parrot] – pájaro [bird]
bebé [baby] – niño [child] (No-error group only)
novia [bride] – mujer [woman] (No-error group only)
pendiente [earring] – joya [piece of jewellery] (Error group only)
nevera [fridge] – frigorífico [refridgerator] (Error group only)

References

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3): 255–278, 2013. ISSN 0749596X. doi: 10.1016/j.jml.2012.11.001.

Štefan Beňuš. Social aspects of entrainment in spoken interaction. *Cognitive Computation*, 6(4): 802–813, 2014. doi: 10.1007/s12559-014-9261-4.

Kirsten Bergmann, Holly P. Branigan, and Stefan Kopp. Exploring the alignment space - lexical and gestural alignment with real and virtual humans. *Frontiers in ICT*, 2(7), 2015. ISSN 2297198X. doi: 10.3389/fict.2015.00007.

Heather Bortfeld and Susan E. Brennan. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23(2):119–147, 1997. ISSN 15326950. doi: 10.1080/01638537709544986.

Richard Y. Bourhis, Howard Giles, Jacques-Phillipe Leyens, and Henri Tajfel. Psycholinguistic distinctiveness: Language divergence in belgium. In *Language and social psychology*, pages 158–185. Basil Blackwell, 1979.

Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25, 2000. doi: 10.1016/s0010-0277(99)00081-5.

Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121(1):41–57, 2011. ISSN 00100277. doi: 10.1016/j.cognition.2011.05.011.

Holly P. Branigan, Alessia Tosi, and Karri Gillespie-Smith. Spontaneous lexical alignment in children with an autistic spectrum disorder and their typically developing peers. *Journal of Experimental Psychology: Learning Memory and Cognition*, 42(11):1821–1831, 2016. ISSN 02787393. doi: 10.1037/xlm0000272.

Susan E. Brennan and Herbert H. Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493, 1996. ISSN 1939-1285. doi: 10.1037/0278-7393.22.6.1482.

Franklin Chang, Gary S. Dell, and Kathryn Bock. Becoming syntactic. *Psychological Review*, 113 (2):234–272, 2006. ISSN 0033295X. doi: 10.1037/0033-295X.113.2.234.

Tanya L. Chartrand and Rick van Baaren. Human mimicry. *Advances in Experimental Social Psychology*, 41(8):219–274, 2009. ISSN 00652601.

Herbert H. Clark. *Using Language*. Cambridge University Press, 1996. doi: 10.1017/cbo9780511620539.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, 1991. doi: 10.1037/10096-006.

Herbert H. Clark and Edward F. Schaefer. Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2(1):19–41, 1987. doi: 10.1080/01690968708406350.

Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1): 1–39, 1986. doi: 10.1016/0010-0277(86)90010-7.

Albert Costa, Martin J. Pickering, and Antonella Sorace. Alignment in second language dialogue. *Language and Cognitive Processes*, 23(4):528–556, 2008. ISSN 01690965. doi: 10.1080/01690960801920545.

Willem Doise, Anne Sinclair, and Richard Y. Bourhis. Evaluation of accent convergence and divergence in cooperative and competitive intergroup situations. *British Journal of Social and Clinical Psychology*, 15(3):247–252, 1976.

Victor S. Ferreira, Daniel Kleinman, Tanya Kraljic, and Yanny Siu. Do priming effects in dialogue reflect partner- or task-based expectations? *Psychonomic Bulletin and Review*, 19(2):309–316, 2012. ISSN 10699384. doi: 10.3758/s13423-011-0191-9.

Alex B. Fine and T. Florian Jaeger. Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, 37(3):578–591, 2013. ISSN 03640213. doi: 10.1111/cogs.12022.

Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. Coming to terms. *Psychological Science*, 23(8):931–939, 2012. doi: 10.1177/0956797612436816.

Simon Garrod and Anthony Anderson. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218, 1987. ISSN 00100277. doi: 10.1016/0010-0277(87)90018-7.

Simon Garrod, Alessia Tosi, and Martin J. Pickering. Alignment during interaction. In *The Oxford Handbook of Psycholinguistics*, pages 575–593. Oxford University Press, 2018. doi: 10.1093/oxfordhb/9780198786825.013.24.

Howard Giles and Peter F. Powesland. *Speech style and social evaluation*. Academic Press, 1975.

Herbert P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*. Brill, 1975. doi: 10.1163/9789004368811_003.

Evelien Heyselaar, Peter Hagoort, and Katrien Segaert. In dialogue with an avatar, language behavior is identical to dialogue with a human partner. *Behavior Research Methods*, 49(1):46–60, 2017. ISSN 15543528. doi: 10.3758/s13428-015-0688-7.

Zoë L. Hopkins and Holly P. Branigan. Children show selectively increased language imitation after experiencing ostracism. *Developmental Psychology*, pages 897–911, 2020. ISSN 00121649. doi: 10.1037/dev0000915.

Zoë L. Hopkins, Nicola Yuill, and Holly P. Branigan. Inhibitory control and lexical alignment in children with an autism spectrum disorder. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 58(10):1155–1165, 2017. ISSN 14697610. doi: 10.1111/jcpp.12792.

William S. Horton and Richard J. Gerrig. Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, 47 (4):589–606, 2002. ISSN 0749596X. doi: 10.1016/S0749-596X(02)00019-0.

William S. Horton and Boaz Keysar. When do speakers take into account common ground? *Cognition*, 59(1):91–117, 1996. ISSN 00100277. doi: 10.1016/0010-0277(96)81418-1.

Ellen A. Isaacs and Herbert H. Clark. References in conversation between experts and novices. *Journal of Experimental Psychology: General*, 116(1):26–37, 1987. ISSN 00963445. doi: 10.1037/0096-3445.116.1.26.

Iva Ivanova and Albert Costa. Does bilingualism hamper lexical access in speech production? *Acta Psychologica*, 127(2):277–288, 2008. ISSN 00016918. doi: 10.1016/j.actpsy.2007.06.003.

T. Florian Jaeger and Neal E. Snider. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83, 2013. ISSN 00100277. doi: 10.1016/j.cognition.2012.10.013.

Anna K. Kuhlen and Susan E. Brennan. Language in dialogue: When confederates might be hazardous to your data. *Psychonomic Bulletin and Review*, 20(1):54–72, 2013. ISSN 10699384. doi: 10.3758/s13423-012-0341-8.

Salla Kurhila. Correction in talk between native and non-native speaker. *Journal of Pragmatics*, 33 (7):1083–1110, 2001. ISSN 03782166. doi: 10.1016/S0378-2166(00)00048-5.

Peter Ludlow. *Living words: Meaning underdetermination and the dynamic lexicon*. Oxford University Press, 2014.

Charles Metzing and Susan E. Brennan. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213, 2003. ISSN 0749-596X. doi: 10.1016/S0749-596X(03)00028-7.

Gary M. Oppenheim, Gary S. Dell, and Myrna F. Schwartz. The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, 114(2):227–252, 2010. doi: 10.1016/j.cognition.2009.09.007.

Jennifer S. Pardo. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393, 2006. ISSN 0001-4966. doi: 10.1121/1.2178720.

Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004. ISSN 0140-525X. doi: 10.1017/s0140525x04000056.

Irina Pivneva, Caroline Palmer, and Debra Titone. Inhibitory control and l2 proficiency modulate bilingual language production: Evidence from spontaneous monologue and dialogue speech. *Frontiers in Psychology*, 3(57), 2012. ISSN 16641078. doi: 10.3389/fpsyg.2012.00057.

David Reitter, Frank Keller, and Johanna D. Moore. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124. Association for Computational Linguistics, 2006. doi: 10.3115/1614049.1614080.

Jorrig Vogels, Emiel Krahmer, and Alfons Maes. How cognitive load influences speakers' choice of referring expressions. *Cognitive Science*, 39(6):1396–1418, 2015. ISSN 15516709. doi: 10.1111/cogs.12205.

Kodi Weatherholtz, Kathryn Campbell-Kibler, and T. Florian Jaeger. Socially-mediated syntactic alignment. *Language Variation and Change*, 26(3):387–420, 2014. ISSN 14698021. doi: 10.1017/S0954394514000155.

Deborah Weiss and James J. Dempsey. Performance of bilingual speakers on the english and spanish versions of the hearing in noise test (hint). *Journal of the American Academy of Audiology*, 19 (1):5–17, 2008. ISSN 10500545. doi: 10.3766/jaaa.19.1.2.

Linda R. Wheeldon and Stephen Monsell. The locus of repetition priming of spoken word production. *The Quarterly Journal of Experimental Psychology Section A*, 44(4):723–761, 1992. doi: 10.1080/14640749208401307.